Minimax linear smoothers

Helge Blaker
University of Oslo

September 1998

Abstract

We consider the problem of estimating the mean of a multivariate distribution. As a general alternative to penalized least squares estimators, we consider minimax estimators for squared error over a restricted parameter space where the restriction is determined by the penalization term. For a quadratic penalty term, the minimax estimator among linear estimators can be found explicitly. It is shown that all symmetric linear smoothers with eigenvalues in the unit interval can be characterized as minimax linear estimators over a certain parameter space where the bias is bounded. The minimax linear estimator depends on smoothing parameters that must be estimated in practice. Using results in Kneip (1994), this can be done using Mallows' C_L-statistic and the resulting adaptive estimator is now asymptotically minimax linear. The minimax estimator is compared to the penalized least squares estimator both in finite samples and asymptotically.

Key words: Nonparametric regression, linear smoother, minimax linear estimator, penalized least squares, Mallows' C_L

1 Introduction

In statistical estimation problems such as curve estimation, signal recovery or image reconstruction, the dimension of the parameter space is of the same order, or even exceeding, the dimension of the data. Methods like least squares or maximum likelihood typically overfit the model, and one needs to restrict the class of possible estimators. One general method for doing this is known as Tikhonov regularization. As in Antoniadis (1996), assume a target function g is observed, where g is a noisy version of a smoother function f and the noise is random. The goal is to reconstruct f. Introduce the error function space $(\mathcal{X}, ||\cdot||_{\mathcal{X}})$ and the smoothness function space $(\mathcal{H}, ||\cdot||_{\mathcal{H}})$ where $\mathcal{H}$ is continuously embedded in $\mathcal{X}$. In this setting, a Tikhonov regularizer of g is $f_k \in \mathcal{H}$ where

$$\inf_{h\in\mathcal{H}}\{||g-h||_{\mathcal{X}}^2 + k||h||_{\mathcal{H}}^2\} = ||g-f_k||_{\mathcal{X}}^2 + k||f_k||_{\mathcal{H}}^2. \tag{1}$$

Here, k is a parameter determining the relative importance of smoothness and fidelity to the data. Equivalently, f_k minimizes $||g-h||_{\mathcal{X}}^2$ over all $h \in \mathcal{H}$ such that $||h||_{\mathcal{H}}^2 \leq \rho$ where ρ determines k. A general alternative to (1) is to find the function f_m such that f_m is measurable relative to g and

$$\inf_{h\in\mathcal{H}} \sup_{f\in\mathcal{H}:||f||_{\mathcal{H}}^2\leq\rho} E||h-f||_{\mathcal{X}}^2 = \sup_{f\in\mathcal{H}:||f||_{\mathcal{H}}^2\leq\rho} E||f_m-f||_{\mathcal{X}}^2 \tag{2}$$

that is, the minimax estimator of f over the parameter space $\Theta = \{f \in \mathcal{H} : ||f||_{\mathcal{H}}^2 \leq \rho\}$. The estimator f_m gives sharp control of the risk $E||h-f||_{\mathcal{X}}^2$ and is more robust than f_k in the sense of having a controlled maximum risk. Of course, it may also take an overly pessimistic point of view.

In principle, this approach can be used generally, but in practice the minimization problem (2) may be much harder than (1). We will therefore restrict attention to the standard model for a random vector $y = (y_1, \ldots, y_n)$ such that

$$y_i = \mu_i + \varepsilon_i, \quad i = 1, \ldots, n, \tag{3}$$

for some $\mu = (\mu_1, \ldots, \mu_n) \in \mathbb{R}^n$ and i.i.d. random variables $\varepsilon_1, \ldots, \varepsilon_n$ such that $E\varepsilon_1 = 0$, $\mathrm{Var}\varepsilon_1 = \sigma^2 < \infty$. The interest focuses on estimating μ. A

special case is nonparametric regression, where one assumes that μ is generated by an underlying function f measured at design points $x_1, \ldots, x_n$, i.e. $\mu_i = f(x_i) = f_i$ where f is assumed to belong to some function class. A regression procedure gives an estimate $\hat{f} = \hat{\mu}$ of μ. This estimate tries to capture the systematic dependency of y on x and is usually smoother than the raw data; methods for estimating f are therefore referred to as smoothing methods. This article concentrates on discrete versions of (1) and (2) where $||g - h||_{\mathcal{X}}^2$ is replaced by $||y - \mu||^2$, $||\cdot||$ is Euclidian norm, and $||h||_{\mathcal{H}}^2$ is replaced by $\mu' B \mu$, B symmetric, non-negative definite. If $\mathcal{X}$ is L^2-space and $\mathcal{H}$ is a q-th order Sobolov space, then the discrete version is obtained from the continuous version using a Fourier expansion and Parseval's theorem. The solution to (1) is then the linear estimator $\hat{\mu} = (I + kB)^{-1} y$ if the inverse exists. Smoothing splines can be put into this framework and minimax estimation of f over linear estimators was considered by Speckman (1985). Carter *et al.* (1992) compare the usual smoothing spline to Speckman's solution.

In general, if $\hat{f} = Sy$ for some matrix S, then $\hat{f}$ is called a linear smoother. If S is symmetric, it is called a symmetric linear smoother. Most commonly used smoothing procedures are linear smoothers, at least if their tuning parameters are considered fixed. Examples are running means, kernel smoothers, regression splines and bin smoothers. Linear smoothers can also be considered basic building blocks in additive models. Buja *et al.*(1989) and Hastie and Tibshirani (1990) consider linear smoothers in great detail. Under order restrictions on the smoothing matrices, Kneip (1994) shows various properties of symmetric linear smoothers taking into account stochastic choice of smoothing parameters (like bandwidth in kernel smoothing or penalty terms for splines). We will restrict the problem (2) further to consider the minimum over linear estimators only. The reason is twofold; firstly, this makes it possible to explicitly solve the minimax problem and compute the minimax estimator; secondly, the results of Kneip (1994) concerning ordered linear smoothers can then be adapted to the minimax estimator to find asymptotically minimax estimators in the more realistic setting when σ^2 and ρ are unknown and must

be estimated from the same data as μ.

Section 2 states and proves a minimax theorem for the minimax problem described above. Buja *et al.* (1989) show that any symmetric linear smoother can be obtained as a solution to a penalized least squares problem. In section 3, we show that symmetric linear smoothers are also minimax linear estimators under suitable restrictions on the parameter space. Section 4 deals with adaptive estimators for the more realistic case when all parameters are unknown and need to be estimated. Finally, some numerical examples are considered.

2 A minimax theorem

Definition 1 *An estimator $\hat{\mu}$ is minimax (relative to the parameter space Θ) if*

$$\inf_{\mu^*} \sup_{\mu \in \Theta} E||\mu^* - \mu||^2 = \sup_{\mu \in \Theta} E||\hat{\mu} - \mu||^2.$$

Here, the inf is over all estimators μ^*. The estimator $\hat{\mu}$ is minimax linear if the inf is over all linear estimators, i.e. of form Cy for some matrix C.

Let $\mathrm{diag}(v)$ where v is a vector be the diagonal matrix with entries v_i on the diagonal. For a symmetric matrix A with eigendecomposition $A = P\Lambda P'$, let $A_+ = P\Lambda_+ P'$ where $\Lambda_+ = \mathrm{diag}(\lambda_i \vee 0)$ which is the positive definite matrix closest to A in trace norm. Let $\mathrm{tr}A$ mean the trace of A, $x_+ = \max(x, 0)$ and δ_{ij} is 1 if $i = j$ and 0 otherwise.

The following theorem can be deduced from results in Pinsker (1980) and Speckman (1985). For completeness we give a self-contained proof.

Theorem 1 *Let $y_i = \mu_i + \varepsilon_i$, $i = 1, \ldots, n$ where $E\varepsilon_i = 0$ and $E\varepsilon_i\varepsilon_j = \sigma^2\delta_{ij}$. Let $\Theta = \{\mu : \mu' B \mu \leq \rho\}$ be the parameter space where B is symmetric, nonnegative definite with spectral decomposition $B = P\Lambda P'$, $PP' = I = P'P$ and Λ is diagonal with elements λ_i being the eigenvalues of B. Let $\mathcal{C}$ be the class of all n by n matrices. Then*

$$\inf_{C \in \mathcal{C}} \sup_{\mu \in \Theta} E||Cy - \mu||^2 = \sigma^2 \sum_{i=1}^{n} (1 - (h\lambda_i)^{1/2})_+ = \sigma^2 \mathrm{tr}(I - h^{1/2} B^{1/2})_+ \qquad (4)$$

where h is determined from

$$\sigma^2 \sum_{i=1}^{n} \lambda_i((h\lambda_i)^{-1/2} - 1)_+ = \rho. \tag{5}$$

The minimax linear estimator is $C^*y = (I - h^{1/2}B^{1/2})_+ y$ and the maximum risk is attained for vectors μ^* such that $\mu^* = P\tilde{\mu}^*$ and $\tilde{\mu}^*$ has components μ_i^* such that $\tilde{\mu}_i^{*2} = \sigma^2((h\lambda_i)^{-1/2} - 1)_+$ if $\lambda_i > 0$ (i.e. $\mu^{*\prime} B\mu^* = \rho$). If $\lambda_j = 0$ for some j, then Θ has no restrictions in the direction of the corresponding eigenvector and $\tilde{\mu}_j^*$ can take any value. If $\lambda_i = 0$ for all i, h is not defined but $\Theta = I\!R^n$ and y is minimax linear.

Proof. Set $\tilde{y} = P'y$, $\tilde{\mu} = P'\mu$, $\tilde{\varepsilon} = P'\varepsilon$ and $\Theta^* = \{\tilde{\mu} : \tilde{\mu}'\Lambda\tilde{\mu} \leq \rho\}$. Then $y = \mu + \varepsilon$, $\mu' B\mu \leq \rho$ transforms to $\tilde{y} = \tilde{\mu} + \tilde{\varepsilon}$, $\tilde{\mu}'\Lambda\tilde{\mu} \leq \rho$ so

$$\inf_C \sup_\Theta E||\mu - Cy||^2 = \inf_C \sup_{\Theta^*} E||\tilde{\mu} - C\tilde{y}||^2 = \inf_C \sup_{\Theta^*} \{||(I-C)\tilde{\mu}||^2 + \sigma^2 \mathrm{tr}(C'C)\}.$$

Set $J(C) = \sup_{\Theta^*} E||\tilde{\mu} - C\tilde{y}||^2$. If C is diagonal with entries c_{ii}, then $J(C) = J_0$ where

$$J_0 = \max_{1 \leq i \leq n} \rho(1 - c_{ii})^2/\lambda_i + \sigma^2 \sum_{i=1}^{n} c_{ii}^2.$$

Let e_i be the ith unit vector. Then

$$\begin{aligned} J(C) &\geq \max_{1 \leq i \leq n} \rho||(I - C)e_i||^2/\lambda_i + \sigma^2 \sum_{i=1}^{n}\sum_{j=1}^{n} c_{ij}^2 \\ &= \max_{1 \leq i \leq n} [\rho(1 - c_{ii})^2 + \rho \sum_{j \neq i} c_{ji}^2]/\lambda_i + \sigma^2 \sum_{i=1}^{n}\sum_{j=1}^{n} c_{ij}^2 \geq J_0 \end{aligned}$$

with equality iff $c_{ij} = 0$, $i \neq j$. The C that minimizes $J(C)$ is therefore diagonal. Let $C = \mathrm{diag}(c_i)$. Since $E(c_i\tilde{y}_i - \tilde{\mu}_i)^2$ is minimized by $c_i = \tilde{\mu}_i^2/(\sigma^2 + \tilde{\mu}_i^2)$,

$$\sup_{\Theta^*} \inf_{(c_i)} E \sum_{i=1}^{n} (c_i\tilde{y}_i - \tilde{\mu}_i)^2 = \sup_{\Theta^*} \sum_{i=1}^{n} \sigma^2\tilde{\mu}_i^2/(\sigma^2 + \tilde{\mu}_i^2) = \nu^2,$$

say. We find ν^2 using the method of Lagrange multipliers. The function

$$g(\tilde{\mu}) = \sum_{i=1}^{n} \sigma^2\tilde{\mu}_i^2/(\sigma^2 + \tilde{\mu}_i^2) + h \sum_{i=1}^{n} \tilde{\mu}_i^2\lambda_i$$

is maximized at $\tilde{\mu}_i^{*2} = \sigma^2((h\lambda_i)^{-1/2} - 1)_+$ where h is determined from $\sum_{i=1}^{n} \lambda_i\mu_i^{*2} = \rho$ and $\nu^2 = g(\tilde{\mu}^*) = \sigma^2 \sum_{i=1}^{n} (1 - (h\lambda_i)^{1/2})_+$. Let $\tilde{c}_i^* = (1 - (h\lambda_i)^{1/2})_+$. Then observe that

$$
\begin{aligned}
\sup_{\Theta^*} E\sum_{i=1}^{n}(\tilde{c}_i^*\tilde{y}_i - \tilde{\mu}_i)^2 &= \sup_{\Theta^*}\sum_{i=1}^{n}\tilde{\mu}_i^2\min(h\lambda_i,1) + \sigma^2\sum_{i=1}^{n}(1-(h\lambda_i)^{1/2})_+^2 \\
&\leq \rho h + \sigma^2\sum_{i=1}^{n}(1-(h\lambda_i)^{1/2})_+^2 \\
&= h\sigma^2\sum_{i=1}^{n}\lambda_i((h\lambda_i)^{-1/2}-1)_+ + \sigma^2\sum_{i=1}^{n}(1-(h\lambda_i)^{1/2})_+^2 \\
&= \sigma^2\sum_{i=1}^{n}(1-(h\lambda_i)^{1/2})_+ = \nu^2 = \sup_{\Theta^*}\inf_{(c_i)} E\sum_{i=1}^{n}(c_i\tilde{y}_i - \tilde{\mu}_i)^2 \\
&\leq \inf_{(c_i)}\sup_{\Theta^*} E\sum_{i=1}^{n}(c_i\tilde{y}_i - \tilde{\mu}_i)^2 \leq \sup_{\Theta^*} E\sum_{i=1}^{n}(\tilde{c}_i^*\tilde{y}_i - \tilde{\mu}_i)^2 \qquad (6)
\end{aligned}
$$

and we have equality throughout in (6), which shows that $\tilde{C}^*\tilde{y}$ is minimax linear. Transforming back to original coordinates, the minimax estimator is $C^*y = P\tilde{C}^*\tilde{y} = P(I - h^{1/2}\Lambda^{1/2})_+P'y = (I - h^{1/2}B^{1/2})_+y$.□

It also follows that C^*y is the Bayes estimator of μ for the problem in which ε and μ are independent normal random vectors, $\varepsilon \sim N(0,\sigma^2 I)$ and $\mu \sim N(0,\Sigma)$, where $\Sigma = \sigma^2(h^{-1/2}B^{-1/2} - I)_+$ and h is determined from $\mathrm{tr}(B\Sigma) = \rho$. More precisely, $E[\mu|y] = \Sigma(\Sigma + \sigma^2 I)^{-1}y = (I - h^{1/2}B^{1/2})_+y = C^*y$.

3 A characterization of symmetric linear smoothers

Consider the penalized least squares problem (where $k \in \mathbb{R}$)

$$
||y - \mu||^2 + k \cdot \mu' A\mu \qquad (7)
$$

where the penalization term $k\mu'A\mu$ depends only on the symmetric part of A, i.e. $\mu'A\mu = \mu'A'\mu$ for all A. Hence only symmetric A will be considered. If inverses exist, the solution is $\hat{\mu} = (I + kA)^{-1}y$ so $S = (I + kA)^{-1}$. Only symmetric smoothers can therefore be obtained by penalized least squares. If S is an arbitrary symmetric matrix with range $\mathcal{R}(S)$, $\hat{\mu} = Sy$ can be characterized as a stationary solution of

$$
Q(\mu) = ||y - \mu||^2 + \mu'(S^- - I)\mu \qquad (8)
$$

under the constraint $\mu \in \mathcal{R}(S)$ as shown in Buja *et al.*(1989) p.466-467. Here, S^- is any generalized inverse such that $SS^-S = S$. If S is nonnegative definite, it can be characterized as a minimizer of (8).

Theorem 1 shows that smoothing matrices corresponding to minimax linear estimators always have eigenvalues in $[0, 1]$. For a characterization of a symmetric linear smoother S as a minimax linear estimator, it is therefore necessary to require that both S and $(I - S)$ are nonnegative definite. If $\hat{\mu} = Sy$ where S has eigenvalues larger than 1, then $\hat{\mu}$ is inadmissible, being dominated by $\{I-(I-S)_+\}y$, where S is replaced by the corresponding matrix where the eigenvalues are truncated at 1.

Theorem 2 *Let T be a symmetric matrix such that T and $(I - T)$ both are nonnegative definite. Then $\hat{\mu} = Ty$ is the minimax linear estimator of μ over the parameter space $\Theta = \{\mu : \mu' B\mu \leq \rho\}$ where $B = (I - T)^2$ and $\rho = \sigma^2\{\mathrm{tr}(T) - \mathrm{tr}(T^2)\}$.*

Proof. Let the eigendecomposition of T be $T = P\Gamma P'$ where $\Gamma = \mathrm{diag}(\gamma_i)$, $0 \leq \gamma_i \leq 1$. Then $B = P(I - \Gamma)^2 P'$ so $\lambda_i = (1 - \gamma_i)^2$ and $\rho = \sigma^2 \sum_{i=1}^n (\gamma_i - \gamma_i^2) = \sigma^2 \sum_{i=1}^n (\sqrt{\lambda_i} - \lambda_i)$, whence theorem 1 shows that $h = 1$ and $Cy = (I - B^{1/2})_+ y = Ty$ is the minimax linear estimator with maximum risk $\sigma^2 \mathrm{tr}(T)$ over Θ. □

Thus any symmetric linear estimator $\hat{\mu} = Ty$ where T has eigenvalues in $[0, 1]$ minimizes

$$\max\{E||\mu^* - \mu||^2; ||(I - T)\mu||^2 \leq \sigma^2(\mathrm{tr}T - \mathrm{tr}T^2)\} \tag{9}$$

over all linear estimators μ^*. The estimator Ty is minimax over the set of parameter values μ where its bias $||E(Ty) - \mu||$ is bounded by $\sigma(\mathrm{tr}T - \mathrm{tr}T^2)^{1/2}$. Alternatively, $E||Ty - \mu||^2 = ||(I - T)\mu||^2 + \sigma^2 \mathrm{tr}(T^2)$ so Θ can be written $\Theta = \{\mu : E||Ty - \mu||^2 \leq \sigma^2 \mathrm{tr}(T)\}$. Notice there are no restrictions on Θ in the directions corresponding to eigenvalues equal to 1.

Example 1. Let T be an orthogonal projection onto a linear space $\mathcal{L}$. Then $T = T^2$ and $\rho = 0$, i.e. T is minimax linear over $\Theta = \{\mu : (I - T)\mu = 0\} =$

$\{\mu : T\mu = \mu\} = \mathcal{L}$.

Example 2. Let $T = \gamma I$ be a constant shrinker, $0 \leq \gamma \leq 1$. Then $\hat{\mu} = Ty$ is minimax linear over

$$\Theta = \{\mu : ||(I - \gamma I)\mu||^2 \leq n\sigma^2(\gamma - \gamma^2)\} = \{\mu : n^{-1}\sum_{i=1}^{n} \mu_i^2 \leq \sigma^2\gamma/(1-\gamma)\}$$

if $\gamma < 1$ and over $\mathbb{R}^n$ if $\gamma = 1$. If $n^{-1}||\mu||^2 \leq a$, then $\gamma = a/(\sigma^2 + a)$ and $\hat{\mu} = a/(\sigma^2 + a)y$ is minimax linear, i.e. if μ^* is any linear estimator,

$$\inf_{\mu^*} \sup_{||\mu||^2 \leq na} n^{-1}E||\mu^* - \mu||^2 = \sup_{||\mu||^2 \leq na} n^{-1}E||\hat{\mu} - \mu||^2 = \sigma^2 a/(\sigma^2 + a). \quad (10)$$

If $\varepsilon \sim N(0, \sigma^2)$, then (10) holds with inf over linear estimators replaced by inf over all estimators as $n \to \infty$, see Pinsker (1980).

Example 3. Let $Sy = (I + A)^{-1}y$ be the solution of (7) when $h = 1$ (i.e. is absorbed into A) and $A = P\Xi P'$, $\Xi = \text{diag}(\xi_i)$, $\xi_i \geq 0$. Let $\tilde{\mu} = P'\mu$ and $B = (I - S)^2$. Then Sy is also minimax linear over

$$\Theta = \{\mu : \mu' B\mu \leq \rho\} = \left\{\tilde{\mu}_i : \sum_{i=1}^{n} \frac{\xi_i^2}{(1+\xi_i)^2}\tilde{\mu}_i^2 \leq \sum_{i=1}^{n} \frac{\xi_i}{(1+\xi_i)^2}\right\}.$$

Conversely, let $T = (I - B^{1/2})_+$, so Ty is minimax over $\Theta = \{\mu : \mu' B\mu \leq \rho/h\}$, where h is determined from (5). Then Ty is also a minimizer of

$$Q(\mu) = ||y - \mu||^2 + \mu'((I - B^{1/2})_+^- - I)\mu.$$

4 Adaptive estimators

The minimax estimator $\hat{\mu} = (I - h^{1/2}B^{1/2})_+ y = C^*(h)y$ over $\Theta = \{\mu : \mu' B\mu \leq \rho\}$ is theoretically determined by the requirement (5), but in practice both σ^2 and ρ will be unknown and hence need to be estimated. This is a general feature of all smoothing problems in practice and popular ways of selecting smoothing parameters include Mallows' C_L, cross-validation or generalized cross-validation. We concentrate on the first. A simple computation shows that for any matrix S,

$$E||Sy - \mu||^2 = E\{||y - Sy||^2 + 2\sigma^2 \text{tr}S - n\sigma^2\}. \quad (11)$$

The expression inside the curly brackets is observable if σ^2 is known and hence gives an unbiased estimate of risk. If $\hat{h}$ minimizes

$$C_L(h) = n^{-1}||y - C^*(h)y||^2 + n^{-1}2\sigma^2 \text{tr}C^*(h)$$

over h, then $C^*(\hat{h})y$ is called a C_L-estimator. This procedure was introduced by Mallows (1973). Kneip (1994) studies large-sample behavior of C_L-estimators for a class of symmetric matrices he calls ordered linear smoothers. Now the class of matrices $\{C^*(h)\}$ where $h \geq 0$ and $C^*(h) = (I - h^{1/2}B^{1/2})_+$ for some nonnegative definite matrix B, is an ordered linear smoother and Kneip's results are applicable. Let the variance estimator be

$$\hat{\sigma}^2 = \frac{y'By}{\text{tr}B}$$

which is always nonnegative and unbiased for σ^2 when $y = \varepsilon$. This is equal to the estimator based on the residual sum of squares $||y - C^*(h)y||^2$ if all eigenvalues of $C^*(h)$ are strictly positive. Assume the following conditions hold as $n \to \infty$.

A There exists a constant $0 < q_1 < \infty$ such that $n\text{tr}(B^2) \leq q_1\{\text{tr}(B)\}^2$.

B There exists a constant $0 < q_2 < \infty$ such that $n^{1/2}\rho \leq q_2\text{tr}(B)$.

Condition **A** makes $\text{Var}\hat{\sigma}^2 = O(n^{-1})$ when $\mu = 0$ while condition **B** bounds the bias $|E\hat{\sigma}^2 - \sigma^2| = O(n^{-1/2})$. Define the normed minimax risk over linear estimators as

$$\nu_n^2 = \inf_C \sup_{\mu \in \Theta} n^{-1}E||Cy - \mu||^2.$$

The following theorem shows that the adaptive estimator $\hat{\mu} = C^*(\hat{h})y$ where $\hat{h}$ minimizes $C_L(h)$ and σ^2 is replaced by $\hat{\sigma}^2$ is asymptotically minimax linear.

Theorem 3 *Assume* $y_i = \mu_i + \varepsilon_i$, $i = 1, \ldots, n$ *where* ε_i *are i.i.d. with* $E\varepsilon_1 = 0$, $\text{Var}\varepsilon_1 = \sigma^2 < \infty$ *and* $E\exp(\beta\varepsilon_1^2) < \infty$ *for some* $\beta > 0$. *Further assume* $\mu \in \Theta = \{\xi \in I\!R^n : \xi' B \xi \leq \rho\}$, *conditions* **A** *and* **B** *hold, and* $\inf_h E||C^*(h)y - \mu||^2 \to \infty$ *when* $n \to \infty$. *Then, as* $n \to \infty$,

$$\sup_{\mu \in \Theta} n^{-1}E||C^*(\hat{h})y - \mu||^2 = \nu_n^2(1 + o(1)).$$

Proof. First assume σ^2 known. Eq. (1.5) in Kneip (1994) (see also comment p.851) gives that under the assumption $E\exp(\beta\varepsilon_1^2) < \infty$,

$$\sup_{\mu\in\Theta}\left(\left(En^{-1}||\mu - C^*(\hat{h})y||^2\right)^{1/2} - \left(\inf_h n^{-1}E||\mu - C^*(h)y||^2\right)^{1/2}\right) \leq dn^{-1/2} \tag{12}$$

for some $d < \infty$. Since $\sup(f(x)-g(x)) \geq \sup f(x) - \sup g(x)$ and $\sup(h(x))^{1/2} = (\sup h(x))^{1/2}$ for $h(x) \geq 0$, (12) implies

$$\left(\sup_{\mu\in\Theta} En^{-1}||\mu - C^*(\hat{h})y||^2\right)^{1/2} - \left(\sup_{\mu\in\Theta}\inf_h En^{-1}||C^*(h)y - \mu||^2\right)^{1/2} \leq dn^{-1/2}.$$

Since $\hat{\mu} = C^*(h)y$ is minimax for some h,

$$\sup_{\mu\in\Theta}\inf_h En^{-1}||C^*(h)y - \mu||^2 = \inf_C \sup_{\mu\in\Theta} En^{-1}||Cy - \mu||^2 = \nu_n^2.$$

As $n \to \infty$, $n\nu_n^2 \geq \inf_h E||C^*(h)y - \mu||^2 \to \infty$ and consequently

$$\sup_{\mu\in\Theta} En^{-1}||C^*(\hat{h})y - \mu||^2 \leq (\nu_n + dn^{-1/2})^2 = \nu_n^2(1 + o(1)).$$

This proves the theorem when σ^2 is known. Theorem 1 in Kneip (1994), on which the above results are based, continues to hold if his eq. (6.1), (6.2) are satisfied. These are exactly conditions **A** and **B** for $\hat{\sigma}^2$ as defined here. Notice $\hat{\sigma}^2$ was chosen so that Kneip's condition $\mu \in V_n(q_2)$ becomes **B** and only relates to the size of Θ, not its shape.□

Remark 1. The condition $\inf_h E||C^*(h)y - \mu||^2 \to \infty$ puts a lower bound on the rate of convergence. In Li (1986), the same condition is used to prove that C_L-estimators are asymptotically optimal for selecting ridge parameters in ridge regression. As he remarks, without this condition resulting estimates may possess unattainably small risk.

Remark 2. If $\varepsilon_1 \sim N(0,\sigma^2)$, $E\exp(\beta\varepsilon_1^2) = (1 - 2\beta/\sigma^2)^{-1/2} < \infty$ for $0 \leq \beta < \sigma^2/2$ and theorem 3 holds. Moreover, the results in Pinsker (1980) show that in the Gaussian case, minimax linear estimators are asymptotically minimax among all estimators and consequently $C^*(\hat{h})y$ is asymptotically minimax. More formally, if the minimax risk over all estimators is

$\delta_n^2 = \inf_{\hat{\mu}} \sup_{\mu \in \Theta} n^{-1} E||\hat{\mu} - \mu||^2$, then $\delta_n^2 = \nu_n^2(1 + o(1))$ as $n \to \infty$ and consequently $\sup_{\mu \in \Theta} n^{-1} E||C^*(\hat{h})y - \mu||^2 = \delta_n^2(1 + o(1))$.

Remark 3. It is clear that some restrictions on B are needed for theorem 3 to hold. If the dimension of B does not approach infinity, then the estimator $\hat{h}$ will not be consistent. For instance, let B be the projection matrix onto the first k coordinate axes. The minimax linear estimator is found to be $\hat{\mu}_i = (1 - k\sigma^2/(\rho + k\sigma^2))y_i$, $\hat{\mu}_i = y_i$, $i \geq k + 1$ so $h^{1/2} = k\sigma^2/(\rho + k\sigma^2)$ while $\hat{\sigma}^2 = y'By/\mathrm{tr}B = \sum_{i=1}^k y_i^2/k$ and $\hat{h}^{1/2} = k\hat{\sigma}^2/\sum_{i=1}^k y_i^2 = 1$. The adaptive estimator of μ_i is thus $\hat{\mu}_i(\hat{h}) = 0$ for $i \leq k$. Condition **A** is satisfied if $k = O(n)$ which again implies that **B** is satisfied if $\rho = O(n^{1/2})$. In that case, $h^{1/2} \to 1$ and the adaptive estimator is consistent.

5 Comparison with penalized least squares

In this section we compare the minimax approach to the penalized least squares method, using both finite-sample calculations and asymptotics. Let the function f have a finite Fourier expansion

$$f(t) = \sum_{j=1}^{n} f_j \phi_j(t)$$

where $\{\phi_j\}_{j=0}^{\infty}$ is a complete orthonormal system on $L^2(Q)$, say. For example, if $Q = [0,1]$ we can take $\phi_0(t) = 1$, $\phi_{2j}(t) = \sin(2\pi jt)/\sqrt{2}$ and $\phi_{2j-1}(t) = \cos(2\pi jt)/\sqrt{2}$. The stochastic process $y(t) = f(t) + \varepsilon(t)$ where $\varepsilon(t)$ is a zero-mean stationary stochastic process with independent increments, can equivalently be represented as

$$y_i = f_i + \varepsilon_i, \quad i = 1, \ldots, n$$

by taking Fourier coefficients ($y_i = (y, \phi_i)$ etc.) and $\int_Q (\hat{f} - f)^2 = \sum(\hat{f}_i - f_i)^2$ by Parseval's theorem. We can therefore consider the discrete version of the problem: estimate f_i when $y_i = f_i + \varepsilon_i$, $i = 1, \ldots, n$ and ε_i are uncorrelated with variance σ^2 under the restriction $\sum_{i=1}^n \sum_{j=1}^n b_{ij} f_i f_j \leq \rho$. If $b_{jj} = j^{2k}$, $b_{ij} = 0$ otherwise, this restriction correspond to the norm of the kth derivative

of f being bounded, but there is no need to restrict the coefficients to have this form.

This usual approach for numerical evaluation of a nonparametric regression procedure is to take a sample of a few 'typical' functions, add some noise and then see how well a procedure reconstructs the original function, see e.g. Antoniadis (1996) sec. 4 or Carter *et al.* (1992) p.88. It is, however, difficult to find a collection of test functions which capture all problematic behavior. The advantage with the Fourier series approach here is that Theorem 1 shows that all that matters is the form of the restricted parameter space so we can focus on the eigenvalues of the matrix B only. Of course, the restrictions on the Fourier coefficients can be translated back to restrictions on the original function.

We will compare the minimax linear estimator to the corresponding penalized least squares estimator for the criterion of maximum expected loss. This can be done explicitly without simulations. Let the restricted parameter space have the form $\sum b_i f_i^2 \leq \rho$ where the restriction to diagonal B is for notational simplicity only since it can always be achieved by a rotation of the coordinate system. Any linear estimator $\hat{f}_i = c_i y_i$ then has maximum risk

$$\begin{aligned} \max_{\sum b_i f_i^2 \leq \rho} E \sum_{i=1}^{n} (\hat{f}_i - f_i)^2 &= \max_{\sum b_i f_i^2 \leq \rho} \{\sigma^2 \sum_{i=1}^{n} c_i^2 + \sum_{i=1}^{n} (1 - c_i)^2 f_i^2\} \\ &= \sigma^2 \sum_{i=1}^{n} c_i^2 + \rho \max_i (1 - c_i)^2 / b_i \end{aligned} \tag{13}$$

where the last equality follows from the proof of Theorem 1. For the minimax linear estimator, this is equal to $R_M = \sigma^2 \sum_{i=1}^{n} (1 - (hb_i)^{1/2})_+$ where h is found from (5). The corresponding penalized least squares estimator minimizes $\sum (y_i - f_i)^2$ subject to $\sum b_i f_i^2 \leq \rho$ and thus has the form $\hat{f}_i = (1 + kb_i)^{-1} y_i$ for some k. Its ideal minimax risk is therefore

$$R_P = \min_k \{\rho k^2 \max_i b_i / (1 + kb_i)^2 + \sigma^2 \sum_{i=1}^{n} (1 + kb_i)^{-2}\} \tag{14}$$

which can be readily computed for known b_i, σ and ρ. This is equivalent to calculations in Carter *et al.* (1992) p.84 but their article only deals with

$b_i = i^4$ and Gaussian errors. We now compute the minimax risk for the minimax estimator and the penalized least squares estimator for a number of different situations. The sample size is either $n = 11$, 41 or 101 and the type of restrictions is either $b_i = i^{1/4}$, $i^{1/2}, i$ or i^2. The relative risk difference $(R_P - R_M)/R_M$ is plotted against ρ for these situations and σ^2 is set to 1. From the form of the estimators, it is clear that the minimax risk increases from 0 for $\rho = 0$ to σ^2 for $\rho = \infty$.

[Figure 1 about here]

This setup reflects different behavior and difficulty in estimating the signal and is of the same type as the one used in Frank and Friedman (1993) in a different context. The parameter $\rho/\sigma^2 = \rho$ when $\sigma^2 = 1$ can be considered the signal to noise ratio. Notice that when b_i is constant, both procedures reduce to $\hat{\mu}_i = \rho/(n\sigma^2+\rho)y_i$ with risk $n\sigma^2\rho/(n\sigma^2+\rho)$. Figure 1 shows that the difference in maximum risk is somewhere between a few percent and quite substantial, especially for small ρ and unfavorable eigenvaluestructure ($b_i = i^d$, $d = 1/4$ or $1/2$) where the risk is increasing in n, especially near 0. These results are confirmed by asymptotic comparison of R_M and R_P.

Theorem 4 *Assume $b_i = i^d/C$ for all i. Then, as $n \to \infty$,*

$$R_M = \left(\frac{\sigma^2 C^{1/d} d}{\rho(d+1)(d+2)}\right)^{d/(d+1)} \rho(d+1)(1+o(1))$$

and

$$R_P = \frac{1}{4}\left(\frac{4\sigma^2 C^{1/d}}{\rho d}\right)^{\frac{d}{d+1}} \rho(d+1)(\Gamma(1/d)\Gamma(2-1/d)/d)^{d/(d+1)}(1+o(1))$$

for $d > 1/2$ and ∞ for $0 < d < 1/2$. Consequently,

$$\frac{R_P}{R_M} \to \frac{1}{4}\left(\frac{(d+1)(d+2)}{d^3}\right)^{\frac{d}{d+1}} (4\Gamma(1/d)\Gamma(2-1/d))^{\frac{d}{d+1}} \tag{15}$$

for $d > 1/2$ and ∞ for $0 < d < 1/2$.

Proof. First consider R_M. Now $\max_i(1-c_i)^2/b_i = \max_i(h \wedge 1/b_i) = h$ for optimal h, otherwise $\hat{\mu} = 0$ with 0 risk. Let $a = h^{1/2}C^{-1/2}$ so

$$\sum_{i=1}^n c_i^2 = \sum_{i=1}^n (1 - h^{1/2} b_i^{1/2})_+^2 \sim \sum_{i=1}^n (1 - a i^{d/2})^2 I(i^d h \le C)$$

$$\sim \int_0^{a^{-2/d}} (1 - ax^{d/2})^2 dx = a^{-2/d} \int_0^1 (1 - y^{d/2})^2 dy = \left(\frac{C}{h}\right)^{1/d} \frac{d^2}{(d+1)(d+2)}$$

Let the risk as a function of h be

$$T_1(h) = \rho h + \frac{\sigma^2 C^{1/d} d^2}{(d+2)(d+1)} h^{-1/d}$$

which has the desired minimum for $h_0 = ((\sigma^2 C^{1/d} d)/(\rho(d+1)(d+2)))^{d/(d+1)}$.

Next consider R_P. Now $c_i = (1+kb_i)^{-2}$ and $\max_i (1-c_i)^2/b_i = k^2 \max_i b_i/(1+kb_i)^2$. Since $\max_x x/(1+kx)^2 = 1/(4k)$, $\max_i (1-c_i)^2/b_i \approx k/4$ for large n. Next, let $a = k/C$ and compute

$$\begin{aligned} \sum_{i=1}^n c_i^2 &= \sum_{i=1}^n (1+kb_i)^{-2} = \sum_{i=1}^n (1+ai^d)^{-2} \sim \int_0^\infty (1+ax^d)^{-2} dx \\ &= a^{-1/d} \int_0^\infty (1+y^d)^{-2} dy = C^{1/d} k^{-1/d} \Gamma(1/d)\Gamma(2-1/d)/d \end{aligned}$$

if $d > 1/2$ and infinity otherwise. Let

$$T_2(k) = \rho k/4 + \sigma^2 C^{1/d} k^{-1/d} \Gamma(1/d)\Gamma(2-1/d)/d$$

which has the stated minimum for $k_0 = ((4\sigma^2 C^{1/d}\Gamma(1/d)\Gamma(2-1/d))/(\rho d^2))^{d/(d+1)}$. □.

Remark 4. The asymptotic risk ratio for $d = 4$ is $(1/4)^{1/5}(45\pi\sqrt{2}/128)^{4/5} = 1.083$ agreeing with calculations in Carter *et al.* (1992). The ratio is strictly decreasing in d, approaching 1 as $d \to \infty$ and approaching infinity as $d \to 1/2$. For the values used in figure 1, the asymptotic values for the relative risk difference are ∞, ∞, 0.225 and 0.115.

The advantage in comparing ideal risks is that no simulations are necessary and we see quite clearly what factors affect the result. For large sample sizes, the adaptive versions of the estimators will have the same behavior when the smoothing parameters are selected using Mallows' C_L. This follows from Theorem 3 for the minimax linear estimator and from Theorem 1 in Li (1986) for the penalized least squares estimator.

Of course, one could also use the adaptive versions of these estimators and compute the observed mean squared errors for the same situations (it is then also necessary to specify the f_i-vector) and the penalized least squares estimator now performs somewhat better than the minimax estimator.

Acknowledgements

This research was supported by grant 121566/410 from the Research Council of Norway.

References

Antoniadis, A. (1996). Smoothing noisy data with tapered coiflets series. *Scand. J. Statist.* **23**, 313-330.

Buja, A.,Hastie, T. and Tibshirani, R. (1989). Linear smoothers and additive models (with discussion). *Ann. Statist.* **17**, 453-555.

Carter, C.K., Eagleson, G.K. and Silverman, B.W. (1992). A comparison of the Reinsch and Speckman splines. *Biometrika* **79**, 81-91.

Frank, I.E. and Friedman, J.H. (1993). A statistical view of some chemometrics regression tools (with discussion). *Technometrics* **35**, 109-148.

Hastie, T. and Tibshirani, R. (1990). *Generalized Additive Models.* Chapman & Hall, London.

Kneip, A. (1994). Ordered linear smoothers. *Ann. Statist.* **22**, 835-866.

Li, K.-C. (1986). Asymptotic optimality of C_L and generalized cross-validation in ridge regression with applications to spline smoothing. *Ann. Statist.* **14**, 1101-1112.

Mallows, C.L. (1973). Some comments on C_P. *Technometrics* **15**, 661-675.

Pinsker, M.S. (1980). Optimal filtration of square-integrable signals in Gaussian noise. *Problems Inform. Transmission* **16**, 120-133.

Speckman, P. (1985). Spline smoothing and optimal rates of convergence in nonparametric regression models. *Ann. Statist.* **13**, 970-983.

Helge Blaker, Department of Mathematics, University of Oslo, P.O.Box 1053 Blindern, N-0316 Oslo, Norway.

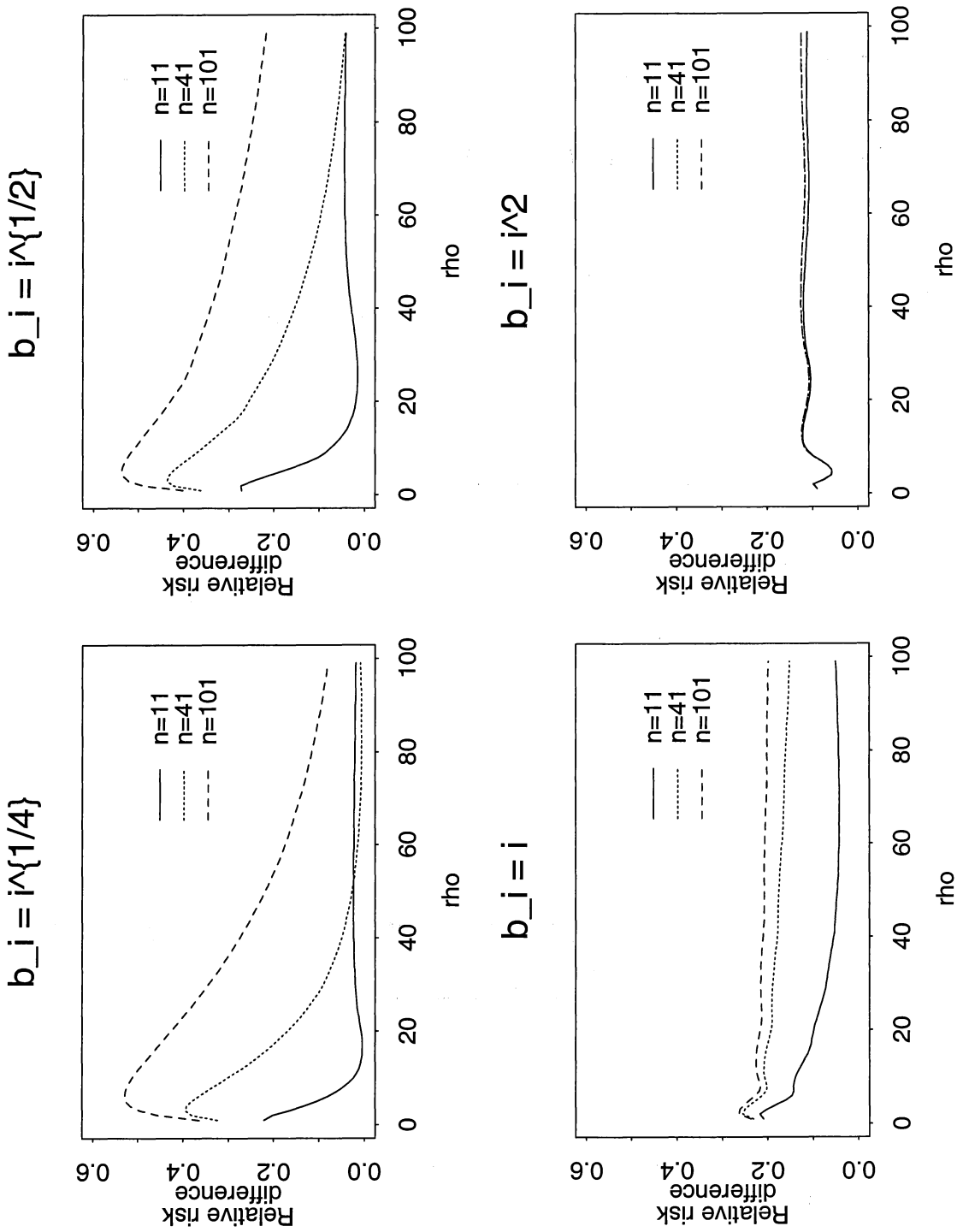

Figure 1: Relative difference in minimax risk